Arrimage avec Learning Resource Metadata Initiative (LRMI)

Marc-Antoine Parent [et Richard Howlett]
GTN-Québec
2015-08-17

Résumé

Parmi ses travaux sur les métadonnées pour les ressources d'enseignement et d'apprentissage (RÉA), le GTN-Québec a misé sur une transition du standard IEEE 1484.12.1–2002 (LOM) vers la norme ISO 19788 (MLR). Nous observons l'émergence de normes concurrentes pour les métadonnées de RÉA, parmi lesquelles, le standard LRMI, soutenu par l'initiative Dublin Core et plusieurs acteurs importants de l'industrie informatique dont Microsoft et Google. Nous avons fait une étude sur les cas d'adoption de cette norme, dans le but de déterminer s'il serait stratégique de développer des outils appropriés pour l'importation, le traitement et l'exportation de métadonnées suivant cette norme. Notre étude montre une adoption limitée, visant surtout l'optimisation pour les moteurs de recherche. Il demeure possible que cet objectif puisse motiver une adoption future du standard et nous recommandons de continuer à observer l'évolution des données LRMI.

Introduction et mise en contexte

Le développement des RÉA vise un objectif simple : permettre aux enseignants et apprenants de trouver des RÉA de qualité en fonction de leurs besoins. Cet objectif prend des couleurs différentes selon les personnes impliquées : producteurs institutionnels ou individuels de ressources libres ou commerciales, développeurs du curriculum, évaluateurs et conservateurs de ressources, enseignants et apprenants. C'est en partie cette diversité, et en partie l'évolution technologique, qui explique l'apparition de plusieurs standards au cours des années.

Énumérons rapidement les principaux acteurs et rôles de l'écosystème des métadonnées :

1. Producteurs des schémas et profils de métadonnées
2. Auteurs de ressources
    - Gouvernementaux
    - Corporatifs
    - Producteurs de contenu indépendants
3. Évaluateurs de ressources

4. Conservateurs de dépôts de métadonnées (y compris les bibliothécaires)
5. Consommateurs de métadonnées (enseignants et apprenants)

Notons en particulier que la production de métadonnées peut venir de chacun de ces acteurs et que l'insistance sera différente selon le cas.

Dans la mesure où les RÉA sont en partie des objets techniques, il est normal qu'un des premiers standards importants, IEEE 1484.12.1–2002 (LOM), soit venu de l'IEEE, en grande partie parrainé par le consortium IMS Global. IMS Global représente en effet des producteurs de RÉA corporatifs dont le contenu technique est important, soit des paquets SCORM, etc. Néanmoins, le standard LOM tient compte de nombreux aspects pédagogiques importants et est conçu selon les normes de bibliothéconomie. Il a permis l'élaboration de dépôts de ressources et de métadonnées, d'ententes de moissonnage, etc. IEEE 1484.12.1–2002 se base sur un modèle abstrait et son expression la plus courante est définie dans IEEE 1484.12.3–2005, basé sur XML.

Le standard LOM permet entre autres de classer des ressources en fonction de taxonomies. Celles-ci sont définies dans un autre standard (VDEX, lui aussi basé sur XML), qui permettent de définir des ensembles cohérents de thèmes, compétences, niveaux pédagogiques, etc. Certaines de ces taxonomies sont inhérentes au standard, alors que d'autres peuvent être étendues dans des profils d'application locaux, comme le profil Normetic du GTN-Québec.

Notons au passage le développement d'autres normes de métadonnées basées sur XML, dont ILOX qui tient mieux compte de certains aspects bibliothéconomiques.

Toutefois, certains biais dans la conception de LOM découlent des éditeurs techniques. Les métadonnées sont regroupées dans une fiche (un seul document XML), et s'il est théoriquement possible pour plusieurs parties (évaluateurs, conservateurs) de publier des fiches de métadonnées différentes autour d'une même ressource, le standard a été conçu pour que chaque ressource ait une fiche canonique. La fiche vient la plupart du temps de l'éditeur de la ressource et les dépôts de métadonnées ne peuvent pas facilement combiner l'information venant de plusieurs sources, dont les évaluateurs et conservateurs.

De plus, la complexité technique relative de XML a freiné la production de métadonnées par des auteurs informels de ressources.

Cette conception d'un document de métadonnées comme provenant d'un seul document n'est pas une fatalité technique. En particulier, le standard RDF, proposé par le W3C avec l'appui de Tim Berners-Lee, permet de considérer ensemble des éléments de métadonnées provenant de sources différentes, ce qui permet une définition plus pluraliste. Il permet également le regroupement des métadonnées selon leur origine. Toutefois, il ne facilite aucunement la composition d'un ensemble cohérent à partir de métadonnées contradictoires.

C'est pourquoi le GTN-Québec contribue au développement de la norme ISO 19788 (MLR), dont l'expression RDF permet aux agrégateurs de données de rassembler des données venant de plusieurs sources, y compris éventuellement des sources informelles. La norme MLR définit en outre une série de vocabulaires taxonomiques propres, définis à l'aide du standard SKOS, ayant des objectifs proches de ceux de VDEX, mais se basant sur RDF. (Notons qu'une expression RDF de LOM a été proposée,

mais elle est peu adoptée.)

RDF propose par ailleurs un modèle distribué, non seulement des métadonnées, mais également des ontologies permettant de définir ces métadonnées. Cette architecture permet à des communautés de définir des formats de métadonnées qui répondent à leurs besoins, en reporte sur les engins d'inférence la difficulté d'interpréter des requêtes pouvant recouper des données définies selon des ontologies disparates. Cette difficulté ne touche pas MLR, qui représente les intérêts d'une communauté, mais touche les moteurs de recherche tels que Google, Yahoo et Bing de Microsoft, qui doivent interpréter cette ontologie parmi plusieurs autres.

Les moteurs de recherche ont réagi en proposant une ontologie unificatrice, Schema.org, dont l'objectif était de donner une ontologie commune pour plusieurs concepts, ce qui les rendrait plus accessibles à un outil de recherche centralisé. Google a profité de sa position sur le WhatWG, un consortium des producteurs de fureteurs qui a forcé la main du W3C pour introduire certaines extensions dans HTML5, pour y incorporer des aspects de Schema.org, ainsi qu'un format limité d'expression RDF dans les documents HTML Microdata. Microdata a vu son rôle diminuer dans les versions définitives de HTML5, mais a néanmoins été bien adopté, car Google favorisait les contenus pourvus de métadonnées suivant les standards Schema.org avec Microdata. Leur incorporation devenait ainsi un élément d'optimisation pour les moteurs de recherche.

L'Association of Educational Publishers, (fondue depuis 2013 avec l'Association of American Publishers), souhaitant à la fois favoriser ses contenus et permettre une mise en évidence des données pédagogiques, a proposé (avec le Dublin Core Metadata Initiative ou DCMI et Creative Commons) en 2012 la première version de LRMI, une extension à Schema.org pour les aspects pédagogiques des RÉA. Schema.org l'a considéré comme appartenant directement à la catégorie des CreativeWork. Sont considérés entre autres les aspects pédagogiques de l'audience, l'âge cible, etc.; le type de RÉA, le type d'interactivité et la notion d'alignement pédagogique. Cette dernière notion permet essentiellement d'associer une ressource pédagogique à une taxonomie pédagogique, sans spécifier la nature technique de ces dernières (sinon qu'il s'agit d'entité RDF ayant présumément une nature propre dans la taxonomie Schema.org) Notons que SKOS ne fait pas parti du standard Schema.org, comme on pourrait s'y attendre.

La recommandation est adoptée par le comité de Schema.org en avril 2013 et la gouvernance de LRMI est confiée à DCMI en octobre 2014. Il s'agit donc d'un standard défini par les producteurs de contenu, avec pour objectif principal d'ajouter des éléments de métadonnées pédagogiques interprétables par les grands moteurs de recherche. Ils pourront les présenter et éventuellement permettre des recherches spécialisées basées sur ces éléments.

Atelier sur l'éducation liée

C'est dans ce contexte que nous avons envoyé Marc-Antoine Parent, du secrétariat technologique du GTN-Québec, comme observateur à l'atelier Linked Education de la conférence ISWC2014. Cet atelier avait explicitement demandé des articles sur LRMI, et nous voulions un meilleur portrait de la pénétration de LRMI dans le mode de l'éducation. Parmi les sept articles, deux mentionnaient LRMI : un article proposant des extensions à LRMI, assez vivement critiqué par les participants et un autre faisant état de l'emploi du Web sémantique sur le site éducatif de la BBC. Il s'agit d'un cas d'école. Ils

ont créé une ontologie pour représenter le curriculum et leurs vidéos pédagogiques sont associés au curriculum par le biais d'Alignements LRMI. Cet alignement leur permet d'employer le curriculum dans une recherche personnalisée Google.

Par ailleurs, dix applications employant les données liées étaient en compétition; la palme a été remportée par Flax, un projet d'apprentissage des langues. Un autre projet très bien reçu était révélateur. Il s'agit de Didactalia, un répertoire de ressources d'apprentissage ouvert en ligne dans la communauté hispanophone. L'équipe de Gnoss emploie très bien les schémas courants du Web sémantique (SKOS, SIOC, DBPedia, etc.), mais est mal informée des standards de ressources pédagogiques, y compris LOM, MLR et LRMI.

En général, l'usage de LRMI semble assez rare; même l'utilisation de schema.org semble réservée à des usages de SEO. Dans la conférence principale (y compris posters etc.), seulement 17 des 226 articles mentionnent Schema.org, et souvent aux fins de recensement ou de critique. La clientèle de la conférence représente un corps de métier autour des données liées et elle est assez critique face à Schema.org. Celui-ci est perçu comme une tentative des moteurs de recherche de se réapproprier le Web sémantique, qui vise la décentralisation. L'auteur d'un article critique détaillé sur la sémantique de Schema.org faisait remarquer qu'un nombre important de joueurs produisent des données dans ce format, mais très peu en consomment en dehors des moteurs de recherche.

État des données LRMI

Voulant en savoir plus sur l'emploi de LRMI dans le Web sémantique, nous avons analysé les données du dernier recensement des données structurées. On y observe une parité approximative entre les données sous format de microdata (utilisant presque toutes Schema.org) et celles en RDFa, qui ne l'emploient que très marginalement. Si on obtient les données microdata elles-mêmes, on trouve que les données relatives aux CreativeWorks représentent un fichier de 6.6Gb (compressé) versus 189 Gb de microdata en général, donc on peut dire qu'elles représentent environ 3.5 % des données. Le fichier lui-même comporte 6 829 740 URIs distincts qui sont des CreativeWork, parmi lesquels 18 977 (0.28 %) ont une propriété learningResourceType et 13 286 (0.19 %) ont une propriété de type educationalAlignment. Les données schema sont également rares dans le répertoire de Mannheim et celui de Cyganiak, sans parler de Sindice.

Par ailleurs, le CETIS a rédigé un rapport d'expériences de cas d'implantation de LRMI. Toutefois, il s'agit d'efforts financés dans cet objectif précis par les fondations Gates et Flora. Il n'est pas clair que ces organisations auraient adopté ce standard sans recevoir ce financement ciblé et il est incertain que cet effort sera poursuivi par ces institutions. Par ailleurs, la diversité d'approches montre que LRMI permet des implantations variées.

Pour aller au-delà de ces études ciblées, nous avons jugé utile de confier un mandat d'exploration à Richard Howlett, un pionnier du Web sémantique à Montréal, afin de lui demander d'évaluer l'adoption de LRMI par des producteurs de RÉA, à l'aide des outils de recherche sémantiques. Ce sont les résultats de cette recherche que nous présentons ici.

Recherche de Richard Howlett

Pour aller au-delà de ces études ciblées, Richard Howlett, un pionnier du Web sémantique à Montréal, évalue l'adoption de LRMI par des producteurs de RÉA, à l'aide des outils de recherche sémantiques. Ce sont les résultats de cette recherche que nous présentons ici LRMI Survey.

Analyse des résultats

Ce recensement, ainsi que les différentes études mentionnées, montre que l'emploi du standard LRMI se répand de façon sporadique chez les producteurs de RÉA, mais ne semble pas encore être devenu un mouvement de fond. Sauf exception, l'objectif dominant semble être celui pour lequel le standard a été conçu, celui d'optimisation des résultats de recherche. Beaucoup d'institutions continuent d'employer leur propre format de métadonnées dans leurs outils internes, pour le traduire en LRMI à l'intention des moteurs de recherche. C'est le cas de la BBC, une des démonstrations les plus complètes de LRMI.

Cet usage n'est aucunement négligeable, et Stephen Dietze, l'organisateur de l'atelier Linked Education, est d'avis qu'il est difficile de faire l'économie de Schema.org. Il serait aisé de transformer nos métadonnées (LOM ou MLR) en données Schema.org. Cette approche tendrait à abandonner aux grands moteurs de recherche le rôle joué par les dépôts de métadonnées. C'est un choix défendable, qui laisserait les institutions pédagogiques se concentrer sur leurs rôles de production de schémas, de classement et d'évaluation.

Toutefois, pour qu'une telle délégation soit souhaitable, plusieurs conditions doivent être remplies :

6. LRMI doit répondre aux principaux critères pédagogiques
7. La production de métadonnées doit être aisée pour les producteurs et évaluateurs de RÉA
8. La recherche selon des critères pédagogiques doit être aisée pour les consommateurs de RÉA
9. Les résultats de recherche doivent être faciles à interpréter

Examinons chacune d'entre elles :

LRMI doit répondre aux principaux critères pédagogiques

Le schéma LRMI convient-il aux besoins de métadonnées pédagogiques? LRMI est un peu moins complet par endroits que le schéma de LOM, mais recouvre largement les éléments obligatoires de Normetic 1.2. En ce sens, il répond aux principaux besoins de classification pédagogique. Un aspect important est que la notion d'alignement est définie de façon suffisamment floue pour qu'il soit possible d'y exprimer une référence aux taxonomies SKOS définies pour MLR, voire aux taxonomies VDEX définies pour LOM moyennant une traduction mécanique vers SKOS.

Le seul élément plus faible qui a trait aux méta-métadonnées, est la notion d'évaluation par un tiers ou d'ensemble de données. Schema.org semble avoir de la difficulté à traiter selon la bonne gestion de LRMI. Nous voyons encore l'influence d'un format défini par les producteurs corporatifs. Cet aspect

mis à part, nous pouvons affirmer que LRMI est un format de données pédagogiques adéquat. Nous pouvons de plus affirmer qu'une partie importante du travail du GTN-Québec, portant sur les taxonomies, serait encore pertinent advenant une décision d'employer LRMI.

La production de métadonnées doit être aisée pour les producteurs et évaluateurs de RÉA.

LRMI n'est pas intrinsèquement plus facile ou plus difficile à utiliser que LOM ou MLR, mais ces formats emploient des langages spécialisés peu connus du grand public. Ils sont peu susceptibles d'être employés sans des outils spécialisés de production de métadonnées, comme le LOMPad du LICEF pour LOM. Certains des partenaires de l'étude du CETIS ont intégré l'édition de métadonnées LRMI à leur chaîne de production de contenu, mais il s'agit d'efforts locaux. Nous n'avons pas encore d'outil d'édition LRMI disponible pour le grand public, et les guides publiés sont destinés aux grands producteurs.

Par contre, l'adoption de LRMI par des dépôts importants tels que le Learning Registry implique que la quantité de données LRMI ira en croissant, et il est plausible que les partenaires du Learning Registry développent des outils pour leur propre usage et qu'ils les partagent entre eux. Il s'agit ici de producteurs corporatifs de RÉA. Rien ne vient faciliter le travail des petits producteurs, évaluateurs ou conservateurs de ressources.

Un autre aspect manquant pour un écosystème en santé est la validation des données LRMI : Google offre des outils de validation pour Schema.org, mais la validation de données propres à LRMI est encore embryonnaire.

La recherche selon des critères pédagogiques doit être aisée pour les consommateurs de RÉA.

L'objectif premier des métadonnées est de permettre la recherche des RÉA par les apprenants et enseignants. Or, en confiant la recherche aux grands moteurs de recherche générique, la communauté de l'enseignement risque de se priver d'une interface permettant une recherche spécialisée sur les critères pédagogiques. Les moteurs de recherches proposent donc la création d'interfaces spécialisées pour les métadonnées de Schema.org. Ce mécanisme a permis à la BBC la création d'un excellent moteur de recherche spécialisé pour ses propres ressources.

Cette solution peut être excellente pour des métadonnées homogènes. Il est assez facile de créer un mécanisme de recherche spécialisé pour les données LRMI, tel que décrit dans cet article. Nous ne pouvons attendre de chaque enseignant qu'il suive une telle procédure, mais il serait en théorie possible de définir une configuration de recherche sur Google accessible à tous.

Toutefois, une telle interface de recherche ne serait généralement valide que pour un ensemble de taxonomies bien défini, ce qui est le cas pour les ressources d'un seul fournisseur ou pour des ressources normalisées. Ni Schema.org, ni ses partenaires n'offrent de mécanisme pour aligner entre elles des taxonomies hétérogènes afin de partager les ressources de différents pays francophones, ce qui a été possible au GTN-Québec avec VDEX et SKOS.

Les résultats de recherche doivent être faciles à interpréter

Un dernier aspect vise l'interprétation des résultats de recherche. Comme mentionné, le modèle de données de Schema.org ne tient pas compte de l'origine d'un ensemble de métadonnées. Il est difficile pour une personne intéressée à une ressource donnée de déterminer si une évaluation positive de la RÉA provient de l'auteur de la ressource, d'un évaluateur, ou d'un autre utilisateur. De plus, ces aspects ne sont pas gérés par les dépôts de données existants, mais le modèle de méta-métadonnées inhérent à LOM et MLR rend cette opération possible.

Conclusion et recommandations

LRMI permet d'identifier une RÉA en tant que telle auprès des grands moteurs de recherche. Dans la mesure où la plupart des consommateurs de RÉA utilisent plus volontiers Google que les dépôts de métadonnées, il est souhaitable d'exposer les métadonnées des dépôts dans le format LRMI. De plus, si ce raisonnement est partagé par la communauté des RÉA, il est plausible que la masse de métadonnées disponibles dans le format LRMI vient à croître avec le temps et qu'il devient souhaitable de pouvoir les moissonner, comme métadonnées LOM ou MLR. La présente étude démontre que nous n'en sommes pas encore à ce stade, mais il convient de surveiller la situation.

Par contre, si l'exportation et éventuellement l'importation de métadonnées LRMI sont souhaitables, nous ne croyons pas qu'il s'agisse d'une raison suffisante pour abandonner les dépôts de métadonnées. Ceux-ci permettent encore des recherches ciblées, les équivalences taxonomiques et surtout une meilleure curation des métadonnées tenant compte de leur origine institutionnelle.

De plus, tant que les outils de production et de validation n'auront pas rejoint le niveau des outils existant pour LOM, il est incertain que l'emploi de LRMI comme format interne pour ces dépôts offre un avantage quelconque sur les formats existants, dans la mesure où il est possible de passer d'une norme à l'autre à l'aide de mécanismes simples (dans la mesure où les vocabulaires peuvent être préservés tels quels.)